基于情感及影响力的微博用户群体特征分析
——以A手机为例

何 跃 尹小佳 朱 超
(四川大学商学院 成都 610064)

摘要:【目的】帮助企业实现精准营销, 准确识别企业用户的群体特征。【方法】对微博文本进行情感分析, 通过 Ward 聚类将微博发表者聚类成 9 类群体, 并对微博用户进行影响力识别, 从情感和影响力两个维度对各个用户群体进行分析, 利用一种改进的客户价值矩阵方法辨别不同用户群体的特征。【结果】实验结果表明: 9 类用户群体对 A 手机品牌情感倾向存在较大的差异。A 手机更受喜欢追赶时髦的女性群体以及从事 IT 行业的用户青睐, 并且该群体影响力较大, 能更有效地影响消费者购买该手机。【局限】在进行用户影响力识别时, 仅考虑常用指标, 未考虑用户微博被转发之后的级联影响力以及其他影响指标。【结论】本文方法能够较为准确地识别企业用户的群体特征, 为企业实现精准营销提供帮助。

关键词: 群体特征分析 情感分析 用户影响力识别 客户价值矩阵
分类号: G353.12
DOI: 10.11925/infotech.2096-3467.2017.0313

1 引 言

Web2.0 时代催生了大量的新型即时通讯工具, 微博就是典型之一, 不仅能满足现代社会大众对信息知晓权的需求, 更充分满足了大众想表达自身意见的欲望。据中国互联网络信息中心(CNNIC)公布的第 39 次中国互联网络发展状况统计报告, 截至 2016 年 12 月, 中国网民规模已达 7.31 亿, 微博用户使用率持续上升达 37.1%[1]。不难发现, 微博用户已经成为中国网民的主要组成部分, 其舆论影响力不可小觑。研究微博平台的用户群体特征, 进而实现大数据精准营销, 已成为热门研究课题。本文选取国内微博平台——新浪微博作为数据收集来源, 以 A 手机为例, 探讨该类手机产品的客户群体特征, 进一步辅助对该产品进行精准营销。

2 研究综述

国内外关于社交媒体上的用户群体特征的研究较多, Li[2]研究了“中国大妈”这类群体使用社交媒体活动的动机和特点, 分析她们对中国社会产生的积极或消极的影响。Koustuv 等[3]收集 Facebook 广告平台的数据, 分析美国各州、性别、年龄、种族亲和度以及教育程度之间的差异。在算法设计方面, Gonzalez- Pardo 等[4]提出一种基于蚁群化优化算法(Ant Colony Optimization, ACO)的改进方法——Bioinspired, 该算法对给定网络中的任一用户, 能够自动确定构成他们兴趣圈的不同用户, 从而进行群体特征识别。Han 等[5]研究用户群体行为对信息传播的影响, 提供一个更好地分析社交网络用户群体行为特征以对信息传播产生影响的参考指标。Step 等[6]使用迭代法进行主题分析, 发现大多数内容主题表达了产品的归属以及相关吸烟活动, 发布最多的用户是男生且年龄较小。

国内近几年, 基于大数据的用户群体特征分析是研究热点, 在新浪微博的研究上, 曾鸿等[7]构建了大数据环境下的用户画像, 选取当代十分具有代表性的

通讯作者: 尹小佳, ORCID: 0000-0002-0850-9720, E-mail: yinxiaojia.xx@foxmail.com。

明星，研究他们的粉丝群体特征。彭希羡等[8]以新浪微博为视角，研究新浪微博上不同的用户性别、认证情况、地区、个性域名、昵称以及描述等指标进行统计分析。在其他社交媒体的用户群体特征研究上，陈梅梅等[9]与淘宝网合作依据全国网络消费者的调查数据，使用消费决策过程理论模型，对比分析我国网络消费者基本属性及购买行为特征，发现我国网络消费者主要关注产品功能、规格和价格，且商品价格承受能力与性别、年龄存在显著关系。符丹等[10]从"海淘族"的用户属性、购物行为和购物体验三个维度出发，分析"海淘族"形成的影响因素及其典型特征。张继东[11]建立了移动社交网络用户行为和偏好的预测机制。

可以看出，国内外对于用户群体特征的分析，不同的学者考虑的角度以及定义的指标不尽相同，而本文则是在结合微博文本情感分析技术、微博用户影响力识别技术以及消费者市场划分技术等的基础上，设计一种适用于微博等社交网络的关于消费者群体的特征分析技术。

目前，国内外学者在用户情感以及用户影响力两方面的研究都较多，但从用户情感和用户影响力两个维度衡量用户群体特征，进而进行精准营销的研究较少。Giatsoglou 等[12]提出一种基于机器学习和文本向量表示的新方法，能快速、灵活地检测出情感片段。Suresh 等[13]采用基于聚类的情感分析方法，提出一种新的模糊聚类模型。在用户影响力研究方面，Jendoubi 等[14]提出基于 Twitter 的两种用户影响力最大化模型，使用信念函数理论估计用户影响。Francalanci 等[15]开发了一种基于 Twitter 网络探索的可视化工具，通过浏览朋友的网络，根据分享内容的实际影响识别关键影响者。Lahuerta-Otero[16]在识别有影响力的用户基础上，分析具有影响力的 Twitter 用户的微博博文内容和数量的特点。这为利用社交网络实施营销提供了新的切入点。

本文以"新浪微博"用户数据为例，采用大数据爬虫技术和机器学习方法对文本进行情感分析，再将微博发表者进行聚类，归纳出不同类别的用户群体。同时采用未确知测度算法对微博群体的不同用户进行影响力识别。最后将各个用户群体在情感倾向和影响力两个不同维度上进行综合分析，利用一种改进的客户价值矩阵的方法辨别出不同用户群体的特征。

3 研究设计

3.1 基于改进词典的情感分析

(1) 数据获取

利用开放源码的网页服务器 Apache2.2，通过新浪微博中提供的 API 接口，在授权的第三方网站上获取数据。收集数据时使用的关键字，主要是关于与 A 手机紧密相连的词组。

①指定采集对象。采集器支持对大量用户批量采集数据，但需要把用户微博地址全部导入至循环列表中。

②设置数据提取字段。根据研究需求，指定网页上需要抓取的数据位置及字段名。

③设置翻页循环。因为主题下的微博数量较多，无法全部显示在一页，所以需要设计翻页循环。

④数据采集。在完成采集流程后，就可以开始采集数据。启动后，采集器会根据设置的流程，对指定的页面网址依次进行采集。

⑤数据导出。在采集任务完成后，可以将采集到的数据导出到 Excel，以便进行数据预处理。

(2) 数据清洗

通过观察收集的数据发现，数据具有随意性、不完整性和多样性等特点，因此，本文数据清洗具体步骤如下：

①通过 Excel 进行数据清洗和预处理，删除乱码数据、不完整数据以及无数据的记录以及原始数据中的垃圾广告；

②采用 volookup 函数进行查找删除，去除步骤①清洗后剩余数据中的相同微博用户，同时去除剩余数据中的相同微博内容的数据；

③使用正则表达式，删除剩余的有效微博文本中包含的网址信息(URL)，只保留文本中的文字、数字、标点和表情符号等信息。将网址去除的正则表达式定义为：

$$http://[\hat{}\backslash u4e00-\backslash u9fa5]^*$$

(3) 基于改进的情感词典的微博情感分析

进行情感分析时，在文献[17]的基础上改进，创新性地提出微博表情识别和转折词的处理两种新的方法。在此基础上进行的情感分析算法不仅极大提高了文本情感识别的准确度，而且在文本处理过于复杂、不易识别的情况下，可通过微博表情识别技术迅速识别出用户情感倾向。

①转折词处理

考虑到转折连词的特点，在处理转折词时，本文提出一种加权的计算转折词前后句情感倾向值的方法。将句子以转折词为界线分为两部分，根据知网情感词典的计算方法分别计算两部分的情感倾向值。最后将两部分分别乘以给定的

权重再相加，得出整个句子的最终情感倾向值。

在确定权重时，根据汉语语法知识，需要确保转折词后部分的情感分值权重大于转折词前的部分[18]。使用 Delphi 专家咨询法通过比较，以此确定句子前后两部分 *fp* 和 *bp* 的权重 α 和 β。处理带有转折词的复句如公式(1)所示。

$$S(sen)=\alpha\times S(fp)+\beta\times S(bp)\quad \alpha<\beta,\alpha+\beta=1 \quad (1)$$

②表情识别

在处理表情符号时，采用内容分析法。内容分析法是一种对研究对象内容进行深入研究和探讨，总结其规律的定性和定量相结合的科学方法。传播学上把它定义为一种系统地、客观地、定量地描绘沟通交流的明显内容的研究方法[19]。

1)确定微博文本中每一个表情符号为一个分析单元；

2)查阅相关文献以及根据内容分析法，制定详细的分析单元归类的标准，确定每一个类目的情感值。在研究中，由于认定表情符号不仅能传递微博发表者较真实的情感，而且还能从不同的表情符号中识别出发表者通过这些表情抒发的情感的强弱。所以在制定类目时，除了要使表情符号能区分情感倾向外，还应赋予每一个表情符号情感的强度。所建立的类目必须满足互斥性、完备性和直观性；

3)邀请 4 位编码员对每一个分析单元做编码，将不同的表情归入不同的类目中。对编码人员进行培训，告知他们本实验的意图以及具体实施方法及步骤。编码人员进行前后 4 轮编码，直到可信度检验结果达到标准为止。最后统计每一个表情符号出现在不同类目中的频数，得到最终的表情符号划分结果；

4)统计每一个表情符号出现在不同类目中的频数，以此对表情符号进行最终的归类。

(4) 用户情感倾向计算流程

微博文本的最终情感倾向值通过上述几种词组和短语的情感值求和得到，当值大于 0 时，表明微博文本所表述的情感为正面情感；当值小于 0 时，所表述情感为负面情感；当值等于 0 时，所表述情感为中立情感。最终的情感倾向值计算如公式(2)所示。

$$S(s)=\alpha\times S(fp)+\beta\times S(bp)+S(e)\quad \alpha<\beta,\alpha+\beta=1 \quad (2)$$

其中，$S(fp)$ 为复句中转折词前的部分情感倾向值，$S(bp)$为转折词后的部分情感倾向值，$S(e)$为表情符号的情感倾向值。

3.2　用户影响力识别

用户影响力识别时，首先运用离差最大化法对评价指标体系进行筛选，再利用分割聚类的方法确定指标评价等级制度[20]，在此基础上计算每一个用户的单指标测度，利用信息熵确定每个用户在每个指标上的权重，再根据单指标测度矩阵以及指标的权重确定综合指标测度矩阵，最后根据置信度准则确定用户影响力等级。

(1) 评价指标体系的选取

本文选取粉丝数、微博被评论数、微博被转发数三个指标[21]，同时创新性地加入了粉丝数/关注数、粉丝数/原创微博数两个指标评价一个用户的影响力。

①粉丝数：该指标表明微博用户被其他用户所关注的程度，是用户影响力最直接的体现。

②微博被评论数：该数值越大说明该用户传达的信息影响范围就越广，该用户的影响力就越大。

③微博被转发数：该数值越大不仅能说明微博用户的微博信息传达给其粉丝这种直接的影响力越大，还包括该信息能被传达给粉丝的粉丝这种间接的影响力。

④粉丝数/关注数：在微博中，有人通过购买粉丝的方式增加自己的粉丝数，提升人气，再与其他微博用户交易，通过大量关注他人的微博来收取佣金。为了在实验中取得较真实的用户影响力值，避免这种毫无意义的“互粉现象”，可以利用“粉丝数/关注数”这一新的变量代替原来的“关注数”指标。

⑤粉丝数/原创微博数：虽然原创微博数标志了一个微博用户在某一话题中的参与度，但是如果发表的微博不能更有效地被其他用户接受，那么这些微博信息将是无效信息。所以，本文提出使用“粉丝数/原创微博数”这一变量代替原来的“原创微博数”。

(2) 评价指标等级标准的划分

本文将用户根据自身影响力划分为：意见领袖、意见活跃分子和普通受众三个等级。使用分割聚类算法中的 K-means 算法将每一个评价指标下关于每个评价对象的值聚为三类，再根据各个类别中的最小值作为等级划分标准的临界值[22]。

该算法在处理数据量较大的聚类时，具有可伸缩性、高效性以及可以同时用于多种数据类型的优点，其算法的时间复杂度上限为：$O(n\times k\times t)$，其中 n 代表对象的数目，t 为迭代的次数。

(3) 基于未确知理论的用户影响力评价模型算法

通过对评价指标体系及其标准的确定后，利用未确知理论算法进行用户影响力值的测算。

①未确知信息测度模型

设 $x_1,x_2,\cdots,x_m$ 为待评价的对象，组成评价对象空间 $X=\{x_1,x_2,\cdots,x_m\}$。对于 $x_i\in X$，有 n 个评价指标 $I_1,I_2,\cdots,I_n$，它们组成评价指标空间 $I=\{I_1,I_2,\cdots,I_n\}$。对于评价对象空间 X 中的每一个评价指标 x，它在某一个评价指标 I_j 下的观测值 x_j 不同时，根据之前确定好的评价等级标准，

将它划入不同的等级标准区间 C_k 的程度λ_{jk}也会不同, 其中 $\lambda_{jk}=\lambda(x_j\in c_k)$。那么可以认为$\lambda_{jk}$是观测值 x_j 使对象 x 处于某种等级标准程度的一种未确知测度, 它必须满足一般测度的三条准则: 归一性、可加性和非负有界性[23]。将测算出的未确知测度写成对于对象 x 的单指标测度矩阵形式[24]如公式(3)所示。

$$(\lambda_{jk})_{n\times p}=\begin{pmatrix}\lambda_{11} & \lambda_{12} & \cdots & \lambda_{1p}\\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ \lambda_{n1} & \lambda_{n2} & \cdots & \lambda_{np}\end{pmatrix} \tag{3}$$

其中, $\lambda_{j1}<\lambda_{j2}<\cdots<\lambda_{jp}$ 或 $\lambda_{j1}>\lambda_{j2}>\cdots>\lambda_{jp}$。

②信息熵确定评价指标系数

观测值x_j使对象x处于某种等级标准程度的未确知测度如公式(4)所示。

$$\lambda_j=(\lambda_{j1},\lambda_{j2},\cdots,\lambda_{jp}) \tag{4}$$

当λ_j中的每个分量取值越集中, w_j取值越大; 当λ_j中的每个分量取值越分散, w_j取值越小。设关于测度λ_{jk}的信息熵[25]如公式(5)所示。

$$H_j=-\sum_{k=1}^{p}\lambda_{jk}\cdot\log\lambda_{jk} \tag{5}$$

根据信息熵理论, 对评价指标 I_j 相对重要程度的不确定性可由熵权值表示[25]如公式(6)所示。

$$V_j=1+\frac{1}{\log p}\sum_{k=1}^{p}\lambda_{jk}\cdot\log\lambda_{jk} \tag{6}$$

③综合测度评价向量

由关于评价对象 x 的单指标测度评价矩阵和权重向量 W, 可求出多指标综合评价测度向量[26]如公式(7)所示。

$$A=W\cdot(\lambda_{np})_{n\times p} \tag{7}$$

④评价准则

由于之前确定好的评价等级是有序的, 故在此不适宜使用最大测度识别准则, 改为使用置信度识别准则。通常设置信度 $\theta(0.5<\theta<1)$ 的值为 0.6 或 0.7, 如公式(8)[26]所示。

$$k'=\min_k\left\{(\sum_{i=1}^{k}\mu_i)\geqslant\lambda,1\leqslant k\leqslant p\right\} \tag{8}$$

则判定评价对象 x_i 属于评价等级 $c_{k'}$。

3.3 用户群体识别

针对于微博及微博中的用户特点, 采用 Ward 聚类方法[27]对目标对象进行聚类分析, 识别出不同特性的用户群体。Ward 聚类方法的基本做法和许多聚类方法一样, 都是先把每个对象看成独立的一类, 逐步归为更大的类。类别每缩减一次, 离差平方和就会增大, 合并使得 S 值增加最小的两个类别, 直到所有的对象都归为一类为止。

3.4 用户群体特征分析

本文针对基于微博的客户细分及用户特征识别问题, 创新性地提出一种改进的客户价值矩阵。使用用户群体平均影响力和用户群体对产品的平均情感倾向值两项指标代替传统客户价值矩阵中平均消费额度以及消费频率两项指标作为分析的两个维度。分别计算出每个用户群体类别的群体平均影响力值 I 和群体总体情感倾向值 S, 再在两个维度建立的矩阵上进行分析。4 个象限的客户类型借鉴“波士顿矩阵”中对 4 种不同产品的命名方法, 将用户群体分为: “明星顾客”, “现金牛顾客”, “问号顾客”以及“瘦狗顾客”[8]。

4 实证分析

数据来源为新浪微博, 实验研究的产品对象为 A 手机。选取在 A 手机上市后两个月的微博数据(2016 年 10 月 28 日–2016 年 12 月 28 日), 共 101 123 条, 涉及微博用户 68 892 名。在对数据进行清洗过滤后, 可用数据剩下 10 853 条, 涉及微博用户 7 043 名。剔除微博内容相同的数据后, 随机选取 5 000 位微博用户关于 A 手机产品的一条微博, 作为实验数据。

4.1 基于改进的情感词典的微博情感分析

(1) 转折词处理

共邀请 20 名专家对句子转折词前后两部分给予赋权。这 20 名专家均是在汉语言文学、数学以及心理学等领域有一定研究成果的专业人士。其中教授 6 名, 副教授 8 名, 在读博士生 6 名; 平均年龄为 42.63 岁, 从事相关研究平均年限为23.45年; 专业方向: 汉语言文学 7 名, 数学 7 名, 心理学 6 名。通过对专家一致性检验, 第 3 轮专家一致性检验结果为 0.82>0.80, 具有较高的可靠性。最终的赋权结果为: 句子转折词前部分权重α为 0.3735, 后半部分权重 β 为 0.6265。

(2) 表情符号处理

从实验数据中总共收集到微博表情 88 个, 本文确定 7 个表情符号类目及其相对应的情感值的表情符号。表 1 为表情符号的最终划分结果。经过 4 轮实验, 在第 4 轮结果的 Kappa 值为 0.81>0.8, 是较好的可信度检验结果, 如表 2 所示。

表 1　表情符号划分最终结果

类目名称	情感值	表情符号
很好	2.5	笑哈哈; 大笑; 嘻嘻; 爱你; 给力; 威武; 顶; 鼓掌; 赞; good; gst 耐你; 好开心
好	2	花心; 可怜; 好激动; 江南 style; 偷笑; 亲亲; 抱抱; 挤眼; ala 加油; 爱心; 耶
较好	1.5	It 切克闹; din 推撞; 兔子; 互粉; 礼物; 微笑; 可爱; 钱; 嘴馋; ok; ala 蹦; 害羞;
稍好	0.5	转发; 围观; 熊猫; 奥特曼; 酷; 猪头; 蜡烛; 坏笑; 勾引
没感觉	0	抠鼻; 浮云; 神马; 时间; 话筒; 疑问; 思考; 国旗;
较差	–1.5	晕; 黑线; 流汗; 囧; 困; 睡觉; 打哈欠; 左哼哼; 右哼哼; 吃惊; 闭嘴; 懒得理你
差	–2	快哭了; 草泥马; xb 压力; 吐血; 衰; 委屈; 吐; 生病; 巨汗; 非常汗; 悲催; 石化; 结冰; 给跪了
很差	–2.5	怒; 怒骂; 抓狂; 崩溃; 哼; 流泪; 鄙视; 失望; 狂躁症; 弱

表 2　可信度检验结果

轮次	Kappa 值
第 1 轮	0.46
第 2 轮	0.59
第 3 轮	0.75
第 4 轮	0.81

(3) 效果对比

经过加入转折词处理以及表情符号处理改进后的词典和传统词典的比较结果如表3所示。由此可以看出，改进后的算法的准确率和召回率的宏平均值都比传统方法高。

表 3　评估结果

评估参数	传统算法得到的结果	改进后的算法得到的结果
Macro-P	0.7362	0.8457
Macro-R	0.7498	0.8590

通过改进后的情感词典的微博情感分析方法，计算出数据集所有微博文本情感倾向值。其中正面情感倾向的微博文本有 3 196 条，中立的微博文本有 26 条，负面情感倾向的微博文本有 1 778 条。可以看出，A 手机在消费群体中的口碑还是比较好的，可以预计手机投入市场之后的销售前景还是比较乐观的。

4.2　用户群体识别

使用 SPSS 17.0 对所有实验样本进行 Ward 聚类分析。采用的聚类指标有 5 个，除了年龄、性别和地域三种个人基本信息以外，还加入了“IT 从业人员、学生及其他”以及“发布微博所使用的终端”两个指标。因为根据市场反应，IT 从业人员和学生为其主要消费者群体。最终将所有微博发布者聚为 9 个不同的类别。通过单因素方差分析表明所有指标的显著性水平(P 值)均小于 0.05，为可接受范围。表 4 为聚类的最终结果以及各个类别的特征以及群体关键字。

表 4　用户群体特征识别结果

群体关键字	用户数目	主要特征
投资者	308	1、主要是金融行业从业者; 大多为男性; 2、主要来自于北京、上海、广东和香港等经济发达地区; 3、微博主要通过 iPhone 手机客户端发布; 4、主要集中在 35-45 岁和 45-55 岁两个年龄段。
IT 业精英	209	1、主要是移动互联网和 IT 企业的企业主和管理层; 2、主要来自于北京和广东两个地区; 3、微博主要通过 iPhone、三星 Galaxy 手机客户端和其他 Android 系统平台发布，其中包含少量小米手机，但比重仅占到 8%; 4、主要集中在 35-45 岁年龄段; 大多为男性。
宅男	465	1、主要集中在 15-25 岁和 25-35 岁两个年龄段; 2、微博主要通过个人电脑或者是类似塞班这样的老式智能手机系统发布。
IT 从业人员	916	1、主要是 IT 企业官方微博和 IT 从业人员; 2、主要来自于北京和广东两个地区; 3、微博主要通过三星 Galaxy，小米手机客户端和其他 Android 系统平台发布，小米手机比重为 33%; 4、主要集中在 25-35 岁和 35-45 岁两个年龄段。

(续表)

群体关键字	用户数目	主要特征
时尚女性	640	1、时尚杂志官方微博，企业白领和主要从事模特、设计师等工作的时尚潮流女士; 2、主要来自于北京、上海、香港和海外; 3、微博主要通过 iPhone 和三星 Galaxy 手机客户端发布; 4、主要集中在 15-25 岁和 25-35 岁两个年龄段。
大龄消费者	378	1、微博主要通过三星 Galaxy、小米手机客户端、塞班和其他 Android 系统平台发布，小米手机比重为 0.02%; 2、年龄段主要集中在 35-45 岁以及 45-55 岁两个年龄段。
智能手机发烧友	552	1、主要是智能手机论坛官方微博以及智能手机分析师、发烧友; 2、主要来自于北京、上海和广东三个地区; 3、主要集中于 25-35 岁年龄段。
宅女	551	1、微博主要通过个人电脑或者是类似于塞班这样的老式智能手机系统发布; 2、主要集中在 15-25 岁和 25-35 岁两个年龄段。
青年学生	981	1、主要集中在 15-25 岁年龄段。

4.3　用户影响力分析

在对微博用户影响力进行分析时，本文选取微博中较为常用的属性指标：粉丝数、微博被评论数、微博被转发数。本文加入两个新的指标："粉丝数/关注数"以及"粉丝数/原创微博数"。表 5 是根据 K-means 聚类算法聚类后的结果。

表 5　各类别评价指标等级标准

等级	1 级	2 级	3 级
粉丝数	$[10000,+\infty)$	$[1000,10000)$	$[0,1000)$
评论数	$[50,+\infty)$	$[1,50)$	0
转发数	$[100,+\infty)$	$[1,100)$	0
粉丝数/关注数	$[100,+\infty)$	$[1,100)$	0
粉丝数/微博数	$[50,+\infty)$	$[2,50)$	$[0,2)$

根据确定好的用户影响力评价指标等级标准以及单指标未确知测度计算方法，构造出关于评价对象的单指标的未确知测度函数。表 6 列举了 6 个用户的影响力识别过程。

表 6　用户各指标数值

用户名	粉丝数(个)	评论数(条)	转发数(条)	粉丝数/关注数	粉丝数/微博数
A	11 305	4	92	25.1222	25.3475
B	42 984	54	200	55.4632	7.4547
C	147 906	0	891	68.3897	7.5824
D	121 846	130	906	74.1607	14.0262
E	1 050	3	7	2.4083	0.2385
F	1 123	4	0	0.5831	2.0912

计算出这 6 个用户的单指标测度值，用矩阵形式表示如下：

$$(\lambda_{1jk})_{5\times3}=\begin{pmatrix}1&0&0\\0&0.0789&0.9211\\0.8384&0.1616&0\\0&0.3260&0.6740\\0&0.9583&0.0417\end{pmatrix}$$

$$(\lambda_{2jk})_{5\times3}=\begin{pmatrix}1&0&0\\0&0.4082&0.5918\\0&0.6054&0.3946\\1&0&0\\0.8309&0.1691&0\end{pmatrix}$$

$$(\lambda_{3jk})_{5\times3}=\begin{pmatrix}1&0&0\\0&0.2041&0.7959\\0.2121&0.7879&0\\0.2975&0.7025&0\\0&0.2326&0.7674\end{pmatrix}$$

$$(\lambda_{4jk})_{5\times3}=\begin{pmatrix}1&0&0\\1&0&0\\1&0&0\\0.4258&0.5742&0\\0&0.5011&0.4989\end{pmatrix}$$

$$(\lambda_{5jk})_{5\times3}=\begin{pmatrix}0 & 0.0111 & 0.9889\\ 0 & 0.0816 & 0.9184\\ 0 & 0.1212 & 0.8788\\ 0 & 0 & 1\\ 0 & 0 & 1\end{pmatrix}$$

$$(\lambda_{6jk})_{5\times3}=\begin{pmatrix}0 & 0.0273 & 0.9727\\ 0 & 0.1224 & 0.8776\\ 0 & 0 & 1\\ 0 & 0 & 1\\ 0 & 0.0038 & 0.9962\end{pmatrix}$$

由公式(5)–公式(7)测算得出以上 6 个用户各个单指标测度评价方案的多指标的综合测度评价矩阵为:

$$A=W\cdot(\lambda_{3\times6})=\begin{pmatrix}0.4190 & 0.3237 & 0.1990\\ 0.6091 & 0.1665 & 0.1544\\ 0.3642 & 0.3469 & 0.2890\\ 0.7829 & 0.1309 & 0.0685\\ 0 & 0.0376 & 0.9624\\ 0 & 0.1166 & 0.9737\end{pmatrix}$$

根据综合测度矩阵 A, 分别对 6 个用户的影响力进行识别并排序, 取 $\lambda=0.6$。在此基础下, 最终得到 6 个用户的影响力如下:

$$p_1=2.1781,\ p_2=2.600,\ p_3=2.0809,$$
$$p_4=2.7880,\ p_5=1.0376,\ p_6=1.0115$$

影响力值排序为 $p_4>p_2>p_1>p_3>p_5>p_6$。

在最终结果中, 用户 C 虽然为 6 个用户里面粉丝数最高的, 但是影响力值并不高, 可以看出, 运用这种综合的未确知测度算法测算出的用户影响力, 能有效排除微博中的僵尸粉、网络水军等的影响, 相较于传统方法判定用户的影响力, 最终结果更客观。

最终, 通过未确知测度模型测算出属于第一影响力等级的微博用户为 47 名, 第二影响力等级的微博发表者为 265 名, 第三等级的微博发表者为 4 688 名。

4.4 结果分析

通过前面的实验数据计算求得最终的基于微博用户群体的价值矩阵, 如图 1 所示。

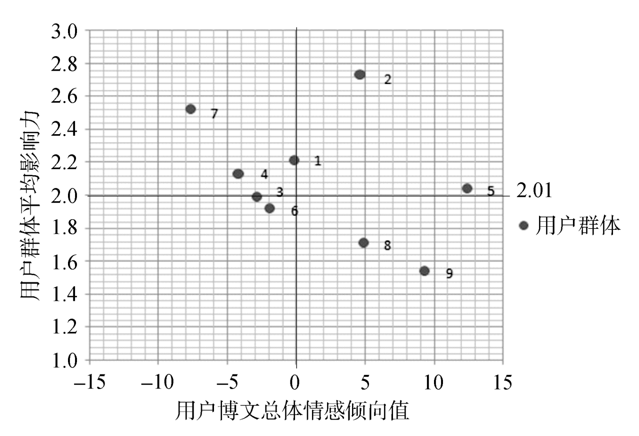


图 1 基于微博的 A 手机客户价值矩阵

从图 1 可以看出, 群体 2(IT 业精英)和群体 5(时尚女性)为"明星用户"这表明, 在 A 手机的所有消费者群体中, 喜欢追赶时髦的女性群体、在 IT 行业从业人员、公司中高层管理者不仅更青睐这项产品, 而且这两个类型的消费者的影响力也较大, 一定程度上能引导其他消费者群体购买 A 手机。因此, 企业在制定营销策略时, 如何为这两类消费者群体提供更优质的产品和服务, 以及如何利用这两类消费者群体创造更高的价值将是关键点。群体 6(大龄消费者)和群体 3(宅男)为"瘦狗"用户。这表明这两类消费者群体对 A 手机不感兴趣或者持负面态度, 在今后的生活中也不大可能购买使用 A 手机。企业在做产品营销时, 对这类用户不应投入过高期望和过多的营销成本和精力。

5 结 语

本文以目前比较流行的智能手机 A 为研究对象, 以新浪微博中收集到的数据为研究样本, 基于微博用户情感分析技术, 用户影响力识别技术以及用户群体特征分析技术进行关于消费者群体特征分析的研究。从实验结果可以看出, A 手机更受喜欢追赶时髦的女性群体以及在 IT 行业从业人员的追捧, 这两个群体能更有效地影响消费者购买该手机, 而大龄消费者和宅男能对 A 手机持否定态度。但文章仍存在一些不足: 在进行文本情感分析时, 新方法效果虽然有进步, 但是并没有考虑网络流行语言以及一些其他因素的影响, 使得有些文本仍然不能被准确识别, 对实验结果造成一定误差。另外在进行用户群体特征分析时, 对于 4 个用户群体象限的划分, 仅仅是根据用户群体平

均影响力值及用户群体总体情感值进行测算，结果并不十分严谨。在今后的研究中，应结合更多影响因子采用更科学的算法进行用户群体象限的划分。

参考文献：

[1] 中国互联网络信息中心. 第39次中国互联网络发展状况统计报告[R/OL]. [2017-01-22]. http://www.cnnic.net.cn/ hlwfzyj/ hlwxzbg/hlwtjbg/201403/t20140305_46240.htm. (China Internet Network Information Center. The 33rd Statistical Report on Internet Development in China [R/OL]. [2017-01-22]. http:// www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201403/t20140305_46240.htm.)

[2] Li Q. Characteristics and Social Impact of the Use of Social Media by Chinese Dama[J]. Telematics and Informatics, 2017, 34(3): 797-810.

[3] Koustuv S, Ingmar W. Characterizing Awareness of Schizophrenia Among Facebook Users by Leveraging Facebook Advertisement Estimates[J]. Journal of Medical Internet Research, 2017,19 (5): e156. DOI: 10.2196/jmir.6815.

[4] Gonzalez-Pardo A, Jung J J, Camacho D. ACO-based Clustering for Ego Network Analysis[J]. Future Generation Computer Systems, 2017, 66: 160-170.

[5] Han S C, Chen H L, Zhang Z J. Influence Model of User Behavior Characteristics on Information Dissemination[J]. International Journal of Computers Communications & Control, 2016, 11(2): 209-223.

[6] Step M M, Bracken C C, Trapl E S, et al. User and Content Characteristics of Public Tweets Referencing Little Cigars[J]. American Journal of Health Behavior, 2016, 40(1): 38-47.

[7] 曾鸿, 吴苏倪. 基于微博的大数据用户画像与精准营销[J]. 现代经济信息, 2016(16): 306-308. (Zeng Hong, Wu Suni. Based on Microblogging Large Data User Portrait and Precise Marketing[J]. Modern Economic Information, 2016(16): 306-308.)

[8] 彭希羡, 朱庆华, 刘璇. 微博客用户特征分析及分类研究——以"新浪微博"为例[J]. 情报科学, 2015, 33(1): 69-75. (Peng Xixian, Zhu Qinghua, Liu Xuan. Research on Behavior Characteristics and Classification of Micro-blog Users—Taking "Sina Micro-blog"as an Example [J]. Information Science, 2015, 33(1): 69-75.)

[9] 陈梅梅, 董平军. 中国网络消费者行为特征[J]. 中国流通经济, 2017, 31(2): 80-85. (Chen Meimei, Dong Pingjun. Behavior Analysis of Chinese Internet Consumer [J]. China Circulation Economics, 2017, 31(2): 80-85.)

[10] 符丹, 刘洪超. "海淘族"的发展与群体特征[J]. 学术探索, 2016(12): 50-55. (Fu Dan, Liu Hongchao. The Development and Group Characteristics of International Shoppers in China [J]. Academic Exploration, 2016(12): 50-55.)

[11] 张继东. 移动社交网络环境下基于情景化偏好的用户行为感知研究[J]. 情报理论与实践, 2017, 40(1): 110-114. (Zhang Jidong. Study on User Behavior Perception Based on Situational Preference in Mobile Social Network Environment [J]. Information Studies: Theory & Application, 2017, 40(1): 110-114.)

[12] Giatsoglou M, Vozalis M G. Sentiment Analysis Leveraging Emotions and Word Embeddings[J]. Expert Systems with Applications, 2017,69: 214-224.

[13] Suresh H, Raj S G. An Unsupervised Fuzzy Clustering Method for Twitter Sentiment Analysis[C]// Proceedings of the 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS). 2016: 80-85.

[14] Jendoubi S, Martin A. Two Evidential Data Based Models for Influence Maximization in Twitter [J]. Knowledge-based Systems, 2017,121: 58-70.

[15] Francalanci C, Hussain A. Influence-based Twitter Browsing with NavigTweet[J]. Information Systems, 2017, 64:119-131.

[16] Lahuerta-Otero E. Looking for the Perfect Tweet. The Use of Data Mining Techniques to Find Influencers on Twitter[J]. Computers in Human Behavior, 2016, 64: 575-583.

[17] 贺飞艳, 何炎祥, 刘楠, 等. 面向微博短文本的细粒度情感特征抽取方法[J]. 北京大学学报: 自然科学版, 2016, 50(1): 48-54. (He Feiyan, He Yanxiang, Liu Nan, et al. A Microblog Short Text Oriented Multi-class Feature Extraction Method of Fine-Grained Sentiment Analysis [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 50(1): 48-54.)

[18] 刘洋. 汉语转折关联词语语义背景探析及教学应用[D]. 济南: 山东大学, 2010. (Liu Yang. The Semantic Backgrounds Study of Adversative Words and Expressions and Application in Chinese Teaching[D]. Ji'nan: Shandong University, 2010.)

[19] Allen B, Reser D. Content Analysis in Library and Information Science Research [J]. Library & Information Science Research, 1990,12(3): 251-262.

[20] 朱郭峰, 杨彦, 周竹荣, 等. 基于领域的微博用户影响力计算方法[J]. 西南大学学报: 自然科学版, 2014, 78(3): 145-151. (Zhu Guofeng, Yang Yan, Zhou Zhurong, et al. Calculation Method of User Influence Based on Domain [J]. Journal of Southwestern University: Natural Science Edition, 2014, 78 (3): 145-151.)

[21] 原福永, 冯静, 符茜落. 微博用户的影响力指数模型[J]. 现代图书情报技术, 2012(6): 60-64. (Yuan Fuyong, Feng Jing, Fu Qianluo. Influence Index Model of Micro-blog

User[J]. New Technology of Library and Information Service, 2012(6): 60-64.)

[22] 冯波, 郝文宁, 陈刚, 等. K-means 算法初始聚类中心选择的优化[J]. 计算机工程与应用, 2013, 49(14): 182-185,192. (Feng Bo, Hao Wenning, Chen Gang, et al. Optimization to K-means Initial Cluster Centers[J]. Computer Engineering and Applications, 2013, 49(14): 182-185, 192.)

[23] 曹庆垒, 李琴, 李丽杰. 基于未确知测度模型的高新区技术创新能力评价研究[J]. 科技管理研究, 2008, 28(5): 134-135. (Cao Qinglei, Li Qin, Li Lijie. Evaluation of Technological Innovation Capability of High-tech Zones Based on Unascertained Measurement Model [J]. Science and Technology Management Research, 2008, 28(5): 134-135.)

[24] 章煜溢, 徐德华. 基于 BSC 和未确知测度理论的 C2C 网商绩效评价模型研究——以淘宝网店铺数据为例[J]. 经营管理者, 2017(4): 4-5. (Zhang Yuyi, Xu Dehua. Study on Performance Evaluation Model of C2C Network Business Based on BSC and Unascertained Measure Theory - Taking Taobao Store Data as an Example [J]. Management Manager, 2017(4): 4-5.)

[25] 周荣虎. 基于信息熵和未确知测度理论的供应链风险系数定量测度模型研究[J]. 中国市场, 2016(45): 52-54. (Zhou Ronghu. Study on Quantitative Model of Supply Chain Risk Coefficient Based on Information Entropy and Unascertained Measure Theory [J]. China Market, 2016(45): 52-54.)

[26] Shanon C E, Weaver W. The Mathematical Theory of Communication [M]. The University of Illinois Press, 1971.

[27] 薛宇, 吴凤平, 王长青, 等. 基于离差最大化和 Ward 系统聚类的医疗服务水平研究[J]. 统计与决策, 2014(16): 86-88. (Xue Yu, Wu Fengping, Wang Changqing, et al. Research on Medical Service Level Based on Maximizing Deviations and Clustering Ward Systems [J]. Statistics and Decision, 2014(16): 86-88.)

作者贡献声明:

何跃: 提出研究方法, 修订论文;
尹小佳: 提出研究思路, 设计研究方案, 撰写论文;
朱超: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: yinxiaojia.xx@foxmail.com。
[1] 尹小佳. 最终数据一览表 1. data.zip. 特征识别表.
[2] 尹小佳. 最终数据一览表 2. data.zip. 微博数据表.
[3] 尹小佳. 最终数据一览表 3. data.zip. 用户信息表.

收稿日期: 2017-04-19
收修改稿日期: 2017-07-31

Analyzing Characteristics of Weibo Users Based on Their Sentiments and Influences —— Case Study of Cell Phone Brand

He Yue Yin Xiaojia Zhu Chao
(Business School, Sichuan University, Chengdu 610064,China)

Abstract: [**Objective**] This study tries to identify the characteristics of consumers, aiming to improve the performance of accurate marketing. [**Methods**] First, we conducted sentiment analysis of the Weibo texts. Then, we divided the Weibo users into nine groups with Ward clustering technique, and identified their influences. Thirdly, we analyzed each user group from the perspectives of sentiment and influence. Finally, we extracted the users' characteristics with a modified customer value matrix. [**Results**] We found significant differences among users' sentiments on a specific cell phone brand. The fashion-chasing women and IT industry workers were in favor of this brand. They could also convince members of other groups choose the same brand. [**Limitations**] We only included the common indicators to examine Weibo users' influences. [**Conclusions**] The proposed method could effectively identify consumers' characteristics and promote accurate marketing.

Keywords: Group Feature Analysis Sentiment Analysis User Influence Identification Customer Value Matrix